COMMENTS on Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research

January 12, 2012

Submitted by Rebecca Kennison, Director, Center for Digital Research and Scholarship, Columbia University (e-mail: rkennison@columbia.edu)

On behalf of Columbia University Libraries/Information Services

New York, New York

Comment 1

The key to encouraging public access to and preservation of digital data is the creation of open data repositories that adhere to common standards for the identification, description, and storage of data, coupled with a policy framework for funding agencies that requires the deposit of data from funded research in such repositories wherever possible, as well as a clear framework for communicating the usage rights for that data.

Given the variability of agency funding (the termination of the National Biological Information Infrastructure [NBII] Program this month is a case in point), the wisest policy is to encourage the growth of existing repositories and the development of new ones that will be managed by individual academic institutions, consortia, and/or scholarly societies in partnership with government, rather than by any individual government agency alone. The National Digital Information Infrastructure and Preservation Program (NDIIPP), which has now become the National Digital Stewardship Alliance (NDSA), offers an excellent example of how agency funding can be used to enhance existing infrastructure investments and encourage inter-institutional (and international) collaboration on data repositories, while the new International Standards Organization standard ISO/DIS 16363, based on the Trusted Repository Audit and Certification (TRAC) checklist (http://public.ccsds.org/publications/archive/652x0m1.pdf), will provide a benchmark against which data repositories will be able to measure their performance. The establishment of baseline metadata requirements for interoperability will also be a key area where agencies can provide leadership, working closely with discipline-specific groups such as professional and scholarly societies, information technology specialists, librarians, and research administrators to ensure that the data in these repositories are stored and described in ways that enhance their discoverability, as well as providing for machine-readability, which will be essential to support the creation of portals that aggregate data from multiple repositories.

Building data repositories and aggregation services alone, however, is not a sufficient step: funding agencies such as the National Institutes of Health (NIH) and the National Science

Foundation (NSF) must move to adopt policies that require researchers to, wherever possible, make their research data available in such repositories to enable public access and reuse. That some agencies have instituted data management plan (DMP) requirements for funding applications has been an important first step in that direction, but further compliance mechanisms will be needed. Those mechanisms should be closely tied to existing workflows for grant management so that important stakeholders, including sponsored projects administrators, repository managers, granting agencies, and the researchers themselves, are minimally burdened by these new requirements, thereby reducing both administrative costs and obstacles to compliance.

These mechanisms will also need to be tied to shifts in the workflows for publication and dissemination of research more broadly speaking. The inability to reproduce and verify research results that serve as the basis for scholarly articles, for example, is a major limiting factor in the efficiency of scientific research, particularly with regard to technology transfer. Since access to research data (and the software that are used to analyze them) is absolutely necessary to ensure reproducibility, there will need to be identification and description standards built into the compliance process that ensure that data are clearly associated with the publications that cite them and the code used to process them. The work of DataCite (http://datacite.org/) offers a promising model in this regard because it makes use of widely-accepted standards such as Digital Object Identifiers (DOIs) to facilitate the discovery, reuse, and impact tracking of data. The work the NIH has done to integrate compliance for published articles based on NIH-funded research into standard publication and grant workflows offers a model for similarly handling data compliance tracking.

Agencies overseas, including the United Kingdom's Joint Information Systems Committee (JISC), the Open Knowledge Foundation, and the Australian National Data Service, also have ample experience in these areas, having undertaken major initiatives in the areas of open data in the past decade. In addition to offering useful models upon which we in the United States can build, their work has begun to demonstrate the wide-ranging and significant economic benefits of open data. A recently commissioned Australian report on public sector information (Houghton 2011: http://ands.org.au/resource/houghton-cost-benefit-study.pdf), for example, found that the benefits for the Australian Bureau of Statistics of moving to Creative Commons licensing for its data would outweigh the costs (including foregone revenues) by 5.3 to 1 and those for the Office of Spatial Data Management and Geoscience by an amazing 15 to 1. A 2008 ACIL Tasman report (http://www.anzlic.org.au/Publications/Industy/251.aspx) suggested that increased public access to spatial data alone brought about a 0.6% to 1.2% boost to Australian GDP and a comparable boost to real wages, as well as a decrease in the trade deficit and an increase in household consumption. Using an econometric methodology, Houghton and Sheehan (2006: http://www.cfses.com/documents/wp23.pdf) determined that a mere 5% increase in access to United States government–funded research results would have produced an additional $2 billion in economic benefits in 2003 alone. Based on 2010 values, that benefit would have been over $7.5 billion, which, using the GDP-to-total-employment ratio, would translate into an additional 700,000 jobs, and, cross-referencing with data from the Bureau of Labor Statistics, a 0.5% decrease in unemployment.

Clearly, then, this is an opportunity that should not be lost: both the path to take and the benefits that will accrue are right in front of us.

Comment 2

The primary challenge here is ensuring that the intellectual property status of data is clearly communicated. For example, works directly produced by the Federal government and its various agencies are part of the public domain as a matter of course, but without their being labeled as such, we have observed that specific uses of the data remain unknown to commons users, thereby thwarting the original intent. Likewise, in the United States, data themselves are not subject to copyright, nor should they be, as any change in their current legal status would impede the technology transfer process and lessen its attendant economic benefits. However, users are frequently unclear as to the disposition of data, a situation that is further complicated by the differing legal frameworks for data in other countries and regions, particularly where data are copyrightable.

Policy in this area, then, should focus on encouraging the clear labeling of data rather than on interferences to the United States' stated position on the noncopyrightable status of data so that all stakeholders — researchers, funding agencies, repository administrators, members of the public, and so on — are aware of the use conditions of a given dataset. That labeling should be done in a human- and a machine-readable format, such as that developed by the Creative Commons (http://creativecommons.org/), with an awareness of the global context in which data live today, as well as of the ever-greater speed with which novel uses for them emerge. It is also vital that government agencies encourage the use of such labeling at all steps of the data lifecycle, since raw data may go through many transformations before they find their way into publications and other end-uses, but the ability to trace those data end-to-end will be an essential part of the verification process.

Comment 3

Federal agencies can expend energy and resources effectively in directing the management of data by embracing and incentivizing cross-institutional partnerships that distribute responsibility and value equally and fairly. The government should certainly assume a role in directing the message about the importance of research data as a first-class scholarly resource, and this would be well expressed in direct and ongoing support for these partnerships — anchored by research institutions with specific domain expertise in concert with the representative scholarly societies, whose membership have the disciplinary focus necessary to identify and address issues of data heterogeneity in the areas of size, sensitivity, format, and long-term value. The library research community, by virtue of its station in cross-institutional information service provision, has been particularly active in assessing the intrinsic issues produced by varying disciplinary data management needs. Federal agency action should therefore be properly informed by

existing studies, including the NSF-backed Data Conservancy project (http://dataconservancy.org/objectives) and the Institute of Museum and Library Services (IMLS)–funded Data Curation Profiles work (http://datacurationprofiles.org/).

Perhaps even more instructive than the differences across disparate disciplinary traditions, however, are the similarities apparent in digital research data that make certain functions common regardless of discipline. The work to be done by federal agencies in promoting and incentivizing best-practice solutions for data archiving and preservation are two such critical vectors. These areas, in particular, may require discrete, focused attention and fostering by federal agencies, as researchers continue to direct limited resources on dissemination goals and problems of access, when they address data management issues at all. A successful plan to address data archiving and preservation will tackle questions of governance, adoption or development of standards and conventions among disciplinary communities, and necessary new investments in technological infrastructure that make data management possible.

Comment 4

At this stage, differentiating between the relative costs and benefits of different data types is premature; the similarities between data types are more significant, and cost–benefit data are limited. What we do know from studies in the United Kingdom and elsewhere is that the costs tend to be localized and front-ended, while benefits accrue broadly and over an extended timespan. Some 42% of costs are associated with pre-archival processing and ingest, while storage and preservation account for 23% and access accounts for 35% (Fry et al. 2008: http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf). Benefits, on the other hand, tend to be wide-ranging and long-term, beginning with the immediate savings produced by avoiding duplicative research and reducing the costs of access to data, followed by the near-term increases in efficiency, and, ultimately, the spillover effects — that is, the broader social and economic gains that result from improved access and that increase over time. Data gathered by the Census Bureau, to take just one US example, are of use to many other government agencies, not to mention a wide range of commercial and non-profit enterprises, and their value is greatly enhanced by their longitudinal nature.

Only by integrating the costs of long-term stewardship and dissemination of data into the granting process will it be possible to gather enough information to allow for a proper consideration of the relative costs of various data types, which will be a prerequisite for an evaluation of the full benefits as well. What is certain, however, is that the overall economic benefits will outweigh the costs. Indeed, as detailed above (see Comment 1), economic studies repeatedly indicate that while the rate of return on investment may vary, the benefits of making data more broadly accessible routinely exceed the costs, even when foregone revenues are taken into account.

Of course, formally integrating these costs into the granting process, while an absolutely vital first step, will not be the end of the story. In addition to setting the stage for further evaluation of the costs and benefits of different data types, agencies must pay attention to the ongoing, unanticipated costs of data stewardship and create mechanisms for meeting those emergent needs that cannot be integrated into and accounted for in the existing grant funding workflows.

One type of data that may be worthy of special consideration from the start is so-called "big data": that is, datasets that are on the order of tera- and petabytes rather than mega- or gigabytes. Due to their enormous size, these data are exceedingly difficult to disseminate via widespread computing networks, even the high-capacity Internet2. In some instances, they are only disseminated on storage media that can be shipped around the country or overseas; in extreme cases, they cannot be disseminated at all, and researchers must visit them directly in order to gain access to them. By adopting policies that encourage the development of functional access solutions for big data, such as migrating large datasets to NoSQL datastores that allow for efficient querying and expanding existing shared computing network infrastructure, funding agencies could have a major impact on the efficiency of scientific research, as well as opening the door to innovative small businesses that cannot afford to run their own high performance computing (HPC) centers.

Comment 5

The better question here is what have proactive stakeholders already been doing to contribute to the implementation of DMPs. Here at Columbia University Libraries and Information Services, for example, we have been working with the Office of Research to coordinate education and outreach efforts on the NIH and NSF DMP requirements, establishing workflows and developing resources to serve the data-archiving needs of Columbia-based researchers, and collaborating directly and indirectly with leading science data groups on campus such as the Center for International Earth Science Information Network (CIESIN) and Integrated Earth Data Applications (IEDA) to ensure that we understand their data preservation needs, collaborations that have led to our partnership with the Socioeconomic Data and Applications Center (SEDAC) on a long-term archive for their data (http://sedac.ciesin.columbia.edu/lta/index.html).

On a national level, there are projects such as those already referenced above: the IMLS-funded Data Curation Profiles project; the NSF-funded DataNet projects, including the Data Conservancy; and the National Information Standards Organization's Institutional Identifier (NISO I^2) project (http://www.niso.org/workrooms/i2). On an international level, we see the emergence of new standards, such as the Open Researcher and Contributor ID (ORCID: http://orcid.org/) and DataCite, that are being developed and supported by a wide range of stakeholders, including researchers, publishers, funding agencies, and research institutions.

The key here is to document the best practices and standards that are emerging and to foster existing systems that already serve research communities. Federal agencies can play an important supporting role in this area in a variety of ways, from funding projects to develop models for data services, to convening meetings to facilitate the codification and dissemination of best practices, to requiring adherence to those best practices and tracking that adherence as part of the grant compliance auditing process.

Comment 6

The first step here is to formally build the funding and tracking of research data costs into the DMP requirements that already exist. As part of that policy shift, agencies will also have to provide greater guidance to applicants to ensure that they are accounting for those costs as best they can and coordinating with the appropriate stakeholders, particularly data repository managers. At many larger research institutions, libraries and other information services providers are already offering what guidance they can on these issues, so there are models that can be built upon; however, guidance coming directly from the funding agencies will carry greater weight with researchers, and it will allow for the development of clear standards that will result in greater uniformity in the resultant cost and benefit data.

Beyond the absolutely vital need to acknowledge the real costs of data management formally by building those costs into research proposal expectations, it is important to recognize that it is not possible to completely account for those costs in advance. In the long term, it is virtually certain that unexpected events will occur that require significant new investment to ensure the ongoing integrity of data. Data migration is a particular challenge that can result in major one-time infrastructure costs: witness the expense of digitizing print media. However, as print digitization projects have made clear, there can also be unexpected benefits to such format migration. The key, then, is for funding agencies to be alert to emerging challenges and opportunities and provide data repositories with the resources they need to meet them.

Comment 7

Automation will be essential to minimizing the burden of compliance. Such automation requires that compliance be integrated into existing workflows (for granting, research, and publication/dissemination) and that it be based on clearly communicated standards. This means that funding agencies will need to improve the instructions they provide to grant writers as they craft their DMPs; it also means that researchers and other stakeholders will agree on basic standards for the identification and description of data, though those baseline standards should remain minimal, with specific disciplines having room to establish their own, more granular standards to meet discipline-specific data and metadata needs.

Comment 8

There are three primary areas in which wider availability of research data will be of short- and long-term benefit to the economy:

- Improving education (both K–12 and postsecondary), to ensure that there is a ready supply of highly skilled individuals ready to enter the STEM workforce;
- Increasing the speed of scientific innovation, to provide for the creation of new technologies and improvements to existing ones;
- Encouraging the growth of small business, by lowering the barriers to entry for high-tech industries and increasing overall competitiveness.

Individual government agencies will have an important role to play in encouraging the benefits in each of these areas, particularly by creating data-aggregating portals that provide a unified point of access to disparately archived data in order to best serve specific stakeholders; as Houghton's models show, the key to increasing the benefits of open data is to provide for the maximum access to that data. Even an increase of 1% can translate into millions of dollars a year. Therefore, efforts like those already undertaken by the Small Business Administration (SBA) to create application programming interfaces (APIs) for other kinds of data should be taken as models for all agencies. Given the SBA's collaboration with NASA's Small Business Innovation Research/Small Business Technology Transfer (SBIR/STTR) group, there are clearly already frameworks in place that would allow for the rapid development and dissemination of tools to maximize the positive economic impact of open data.

The Data.gov portal is another important platform for emerging and new markets. Because of its role as a clearinghouse for open data, it is especially useful for attracting interest from application developers, and thus for generating new technologies and new businesses that can provide additional services to the public. Thus, integrating open research data into Data.gov would offer greater potential for spillover benefits, particularly from unforeseen uses of research data, including but certainly not limited to data mash-ups, and innovative (and perhaps serendipitous) discovery resulting in new products.

Comment 9

There are at least two such mechanism-types that could be brought to bear on the attribution of data for secondary research: (1) the improvement and disciplinary standardization around persistent identifiers for data, institution, and researcher; and (2) the incentivizing of the practice of data citation in a way that raises it to the scholarly standard of publication citation already commonly well-observed. To the former, the issue of identifier persistence is one that has been taken up by several parallel groups (ORCID for researchers, NISO I^2 for institutions, and DataCite for the data themselves). Through identifier persistence, explicit stakeholder credit may be tied directly to unique data assets, which strengthens not only the potential of explicit attribution but also for the potential to trace reuse through citation, a significant value proposition to researchers looking for the

rationale behind data sharing. Rather than look for new mechanisms to begin developing, therefore, a powerful way to support existing momentum would be direction for federal agencies to provide incentives around their use — much in the model of the NIH in its program to ensure funded research publications in the PubMed Central repository are also cataloged within the freely accessible PubMed database. Further, an early driver of potential movement could come from federal agencies to catalyze research around publicly accessible datasets, with specific caveats about the methods of attribution to be adhered to. Such programs would reward early adopters of best practices in data management while contributing to the hoped-for sea change in source acknowledgement in secondary research.

Comments 10 and 11

International standards efforts, such as the recently approved standard ISO 16363 (Space Data and Information Transfer Systems — Audit and Certification of Trustworthy Digital Repositories) and the development of the proposed standard ISO 16919 (Space Data and Information Transfer Systems — Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories), offer one approach to improving the likelihood of success of standards development processes. Both efforts were conducted under the guidance of the Consultative Committee for Space Data Systems (CCSDS) by a diverse, international team of experts in archiving and data stewardship. Team members developed their recommendations independent of influence from government agencies or proprietary interests, which enabled the team to address key standards issues without being compromised by other influences. The development of these standards was encouraged and supported by the institutions that employed the team members and by the funders of those institutions. Such support and encouragement enabled the experts to collaborate as a cohesive team over extended periods of time, ensuring that the process had continuity in terms of the stakeholders and the expertise brought to bear.

Many of the standardization efforts around attribution and credit for data reuse (noted in the response to the previous question) will apply here as well in the context of interoperability, and our suggestions regarding the support of those standards should be carried over. Further, there are many established disciplinary data and metadata standards and communities of which of the RFI reviewers will be well aware — the Flexible Image Transport System (FITS) for astronomical data, and the Federal Geographic Data Committee (FGDC) and the Open Geospatial Consortium (OGC) for geospatial data, are several that have special relevance and utility to the Columbia University community. The multiplicity of these standards prohibits a comprehensive response here, although we do join in the endorsement for a centralized index of such standards, believing that such a resource could foster adherence to community practice and reduce barriers to interoperability.

Perhaps more significantly, however, we invoke once more the call to involve the scholarly and professional societies in a direct way in the identification and development of these domain-specific digital data standards and of the data repositories themselves. As both liaisons among and representatives for their constituencies, societies are equipped to deal with the inevitable idiosyncrasies of the data in their domain. Empowering these organizations (again, through incentives articulated centrally through individual agencies) thus strengthens their positions as arbiters of authority and respects the individual established contexts, initiatives, and standards.

Comment 12

Given that such initiatives to coordinate on data standards are presently underway through several international bodies, including the ISO and the International Council for Science (ICSU) and its Committee on Data for Science and Technology (CODATA), we can advise in the first place for federal support for the activities of these groups and others of a disciplinary orientation as they represent national interest in international consensus. Further, the direct involvement of the National Institute of Standards and Technology (NIST) in matters of standards-making with our global counterparts would be a natural activity, and we can advise for further federal agency involvement in related activities. This can be understood to mean a push for proactive engagement of national agencies with initiatives of nations, some such as the United Kingdom whose work on data standards is tracking very closely with activity in the United States. The relevant National Research Council boards may also play a constructive role in this arena, e.g., the Board on International Scientific Organizations (BISO) and the Board on Research Data and Information (BRDI).

Comment 13

For the appropriate association between research article and dataset to be made, it is not enough that digital research data be attached as supplementary material to published research articles. It is, in fact, the research article that is the supporting documentation of primary research data, although the infrastructure around the publication and archiving of written material is much more mature than that around datasets. The research article itself may come to be supplemented by secondary research against the original dataset or only a subset of it, but in any robust future scenario, the dataset needs to occupy a position of importance above that of supplementary material. The work advocated in the responses to the previous questions — particularly in the support of archiving and preservation environments for datasets, as well as for community coalescence around persistent identifier schema — are important parts of the publication/data association infrastructure. Look to the German academic institution-based Publishing Network for Geoscientific and Environmental Data (PANGAEA, a DataCite member: http://www.pangaea.de) and its formal affiliations with Elsevier for an example of how linkages and partnerships may begin to be realized in practice (http://bit.ly/9SSkHQ). The work of the international data

repository called Dryad in facilitating data and publication links in additional to data reuse is also particularly instructive (http://datadryad.org/).